From Simulation Data to Conformational Ensembles: Structure and Dynamics-Based Methods

WILHELM HUISINGA,[1] CHRISTOPH BEST,[2] RAINER ROITZSCH,[1] CHRISTOF SCHÜTTE,[1] FRANK CORDES[1]

[1]*Konrad–Zuse–Zentrum Berlin, Takustraße 7, D-14195 Berlin, Germany*
[2]*Forschungszentrum Jülich (NIC), Germany*

Received 4 May 1999; accepted 18 August 1999

ABSTRACT: Statistical methods for analyzing large data sets of molecular configurations within the chemical concept of molecular conformations are described. The strategies are based on dependencies between configurations of a molecular ensemble; the article concentrates on dependencies induced by (a) correlations between the molecular degrees of freedom, (b) geometrical similarities of configurations, and (c) dynamical relations between subsets of configurations. The statistical technique realizing aspect (a) is based on an approach suggested by Amadei et al. (Proteins 1993, 17). It allows identification of essential degrees of freedom of a molecular system, and is extended to determine single configurations as representatives for the crucial features related to these essential degrees of freedom. Aspects (b) and (c) are based on statistical cluster methods. They lead to a decomposition of the available simulation data into conformational ensembles or subsets, with the property that all configurations in one of these subsets share a common chemical property. In contrast to the restriction to single representative conformations, conformational ensembles include information about, for examples, structural flexibility or dynamical connectivity. The conceptual similarities and differences of the three approaches are discussed in detail, and are illustrated by application to simulation data originating from a hybrid Monte Carlo sampling of a triribonucleotide.
 J Comput Chem 20: 1760–1774, 1999

Keywords: conformational ensemble; cluster method; structural and dynamical similarity; essential degrees of freedom; feature extraction

Correspondence to: W. Huisinga; e-mail: huisinga@zib.de
Contract/grant sponsor: Deutsche Forschungsgemeinschaft (DFG); contract/grant number: De 293

CCC 0192-8651 / 99 / 161760-15

Introduction

A molecule can exist in an infinite number of spatial states, in literature also known as configurations. Subsets of states can be identified as conformational ensembles, if all states within the subset share a common chemical property. However, often just a single state is refered to as a conformation, being a typical representative of the chemical property. In this article, we consider both aspects with the main emphasis lying on the ensemble aspect of conformations.

Structurally *meta*-stable conformations belong to minima of the free-energy landscape, which are expected to be nearly isoenergetic.[1] They could be identified experimentally, if the barriers between the minima enabled lifetimes of the states, which correspond to the time resolution of the experiment. Transitions between these conformations have been investigated theoretically; they can be described by a few collective interdomain motions, which correspond to low energy conformational changes.[2, 3] Functionally active conformations can be characterized by their ability to initiate a process. They need neither to be structurally conserved in every atom position nor necessarily occupy an energy minimum. Up to now, the lifetime of these conformations is generally below experimental resolution, but within the scope of simulations.

Theoretical investigations of simulation data have contributed to the understanding of, for example, conformational gating of buried active sites,[4] binding of ligands to proteins,[5] folding processes of proteins,[6, 7] or the initial state of an enzymatic reaction.[8] These insights result from analyzing simulated time series below the nanosecond range, which are assumed to correspond to thermodynamical ensembles of the respective molecular system. Although the simulation data are still not large enough to accurately reproduce configurational entropy or free energy, the classification in terms of conformational subsets may help to understand the structure–function relationship of biomolecules and allows reduction of the complexity without losing relevant information.

Cluster methods are well established in many fields to discover the unknown structure of complex data.[9, 10] In our context, these methods can be used to decompose some large set of configuration data into disjoint conformational subsets with the property that configurations from one subset are structurally closer to each other than configurations from different subsets. There are two classes of such structure-based cluster methods: (1) Global methods directly handle the entire set of data; they are known from, for example, clustering in large graphs via spectral information[11, 12] or from decomposition techniques using multiple-center estimations of phase space distributions.[13] (2) Sequential methods handle the data set as a sequence and refine the decomposition iteratively; they are based on, for example, nonhierarchical neural nets[14] or fuzzy clustering.[15] However, none of the structure-based cluster methods takes into account any dynamical properties of the molecular system underlying the given set of configurations. That is, these methods do not allow to characterize the *meta*-stability of the conformational subsets, for example, their lifetimes or the rate of transitions between them. Dynamics-based cluster techniques have been developed only recently.[16] Based on additional transition data, the set of configurations is decomposed into conformational subsets with respect to *meta*-stability minimizing the transition probabilities between them.

In ref. 17, Amadei et al. introduced the so-called essential dynamics techniques tailored to identify essential degrees of freedom of molecular systems based on simulation data. We will demonstrate how this technique can also be used as a tool for conformational analysis (see Representative Conformations in the Methods section). In contrast to the cluster analysis tools discussed above, this approach yields single representative conformations rather than conformational subsets or ensembles.

The primary concern of this article is to discuss and compare the structure-based and dynamical-based conformational subsets. To this end, two different cluster techniques and the representative analysis are considered in detail, at first conceptionally (see Clustering Methods in the Methods section), and then via application to the triribonucleotide adenylyl(3′-5′)cytidylyl(3′-5′) cytidin [r(ACC)] in vacuum, chosen because of its structural flexibility (see Results).

The required configuration and transition data were computed by means of a specific hybrid Monte Carlo method with adaptive temperature choice (ATHMC)[18] especially designed to overcome energy barriers. Based on the resulting data, we observe some interesting and far-reaching similarities between dynamics and structure-based conformational subsets but also significant differences, for example, situations in which configurations from different *meta*-stable conformational subsets are clustered together by the structure-based method.

Methods

In classical MD, a molecule is modeled by a Hamiltonian function

$$H(q,p) = \frac{1}{2} p^T M^{-1} p + V(q),$$

where $q = (q_1, \ldots, q_{3N})$ and $p = (p_1, \ldots, p_{3N})$ are the corresponding positions and momenta of the atoms, M the diagonal mass matrix, and V a differentiable potential. The formal solution with initial state $q_0 = q(0)$, $p_0 = p(0)$ is given by $q(t) = q(t; q_0, p_0)$ and $p(t) = p(t; q_0, p_0)$.

Most experiments on molecular systems are performed under the conditions of constant temperature and volume. The distribution of the identically prepared systems is described by the stationary canonical density associated with the Hamiltonian H

$$f(q,p) = \frac{1}{Z} \exp(-\beta H(q,p)),$$

where Z denotes the partition sum, $\beta = 1/k_B T$, with T being the system's temperature T and k_B Boltzmann's constant. Because H is separable, the canonical density decomposes into a product of a position density $Q(q)$ depending only on q, and a momentum density $P(p)$ depending only on p, i.e., $f(q,p) = Q(q)P(p)$.

The starting point for the methods presented below is a sampling of the positional part of the canonical density resulting in simulation data $q^{(1)}, \ldots, q^{(S)}$. When using reweighted hybrid Monte Carlo methods, there is a weighting factor associated to each state $q^{(k)}$. To keep the notation simple, we assume within this section the weighting factor being the same for all states.

REPRESENTATIVE CONFORMATIONS

Identifying representative conformations is based on a correlation analysis of the molecular degrees of freedom. Analysis of simulation data reveals that it is possible to divide the set of molecular degrees of freedom into two subsets: a subset of only a few "essential" degrees of freedom, in which anharmonic motion occurs that comprises most of the positional fluctuations, and the remaining degrees of freedom, in which the motion has a narrow Gaussian distribution and can be considered as "physically constrained."[17] This is in contrast to normal mode analysis,[19] which is based on the shape of the potential energy function and is restricted to predict only harmonic vibrations. We determine essential degrees of freedom both in the Cartesian coordinate space following Amadei et al.,[17] as well as in the space of torsion angles.

Cartesian Coordinates

Because essential degrees of freedom should only reflect internal fluctuations of the molecule, we first eliminate the overall translational and rotational motion. This is done by a least-squares translational and rotational fit of each configuration to an arbitrary chosen reference configuration. The results are fitted simulation data, which, for simplicity, we again denote by $q^{(1)}, \ldots, q^{(S)}$.

The correlation between atomic motions within the simulation data are expressed by the covariance matrix

$$C = \mathrm{Cov}(q_k, q_n)_{k,n=1,\ldots,3N} = \left\langle (q - \langle q \rangle_Q)(q - \langle q \rangle_Q)^T \right\rangle_Q,$$

where $\langle \cdot \rangle_Q$ denotes the ensemble average, i.e.,

$$\langle q \rangle_Q = \int_\Omega q Q(q)\, dq \approx \frac{1}{S} \sum_{s=1}^{S} q^{(s)}.$$

Because C is symmetric, it can always be diagonalized. Let U denote the matrix, whose columns are the eigenvectors of C. Then the transformed coordinates $x = U^T(q - \langle q \rangle_Q)$ are uncorrelated and

$$\Lambda = \mathrm{Cov}(x_k, x_n)_{k,n=1,\ldots,3N} = \left\langle x x^T \right\rangle_Q$$

for a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_k)$. If we choose the eigenvectors to be normalized with respect to the Euclidean norm (or equivalently, the matrix U to be orthogonal), the eigenvalues are equal to the variance of the transformed coordinates, i.e., $\mathrm{Var}(x_k) = \lambda_k$ for all k.

Due to the last equality, the covariance matrix is connected to the systems constraints in the following way:[17] transformed coordinates corresponding to zero or nearly zero eigenvalues behave effectively as constraints; they have narrow Gaussian distributions with a zero mean, and do not contribute significantly to the positional fluctuations. In contrast to that, transformed coordinates corresponding to large eigenvalues represent large positional deviations. Often, only a few coordinates see important fluctuations; these are called essential degrees of freedom. In practice, one has to specify a set of largest eigenvalues of C, which often can only be done heuristically.

Torsion Angles

An alternative way to describe conformational changes is based on the set of torsion angles

$\omega_1, \ldots, \omega_M$ of the molecular system under consideration. To analyze the simulation data in terms of these torsion angles we have to apply statistical methods for circular data.[20, 21]

The mean direction $\mu(\omega_k)$ of the torsion angle ω_k and a corresponding circular deviation $\rho(\omega_k)$ are defined by

$$r(\omega_k)\exp(i\mu(\omega_k)) = \langle \exp(i\omega_k)\rangle_Q \quad \text{and}$$
$$\rho(\omega_k) = \sqrt{-2\log(r(\omega_k))}.$$

The value $r(\omega_k)$ is called the mean resultant length. To apply the analysis from above, we choose the following definition of correlation between circular variables[20, 21]

$$\mathrm{Cor}(\omega_k, \omega_m) = \frac{r^2(\omega_k - \omega_m) - r^2(\omega_k + \omega_m)}{\sqrt{(1 - r^2(2\omega_k))(1 - r^2(2\omega_m))}}$$

and set $\mathrm{Cov}(\omega_k, \omega_m) = \mathrm{Cor}(\omega_k, \omega_m)\rho(\omega_k)\rho(\omega_m)$. These definitions allow to apply the technique introduced above for simulation data $\omega^{(1)}, \ldots, \omega^{(S)}$ being converted to torsion angles. [In contrast to the Cartesian coordinate case, there are different definitions of correlations between circular data (see refs. 28 and 30, and references therein).]

The identification of representative conformations is based on the following idea. According to their distribution, essential degrees of freedom can roughly be divided into two groups: (1) broad Gaussian shaped and (2) multiple peaked compounding of Gaussian-like parts. Each Gaussian peak corresponds to a part of the configurational space with a relevant weighting factor, and might be represented by a configuration associated with the maximum of the Gaussian peak. Thus, configurations associated with a combination of maxima of all essential degrees of freedom correspond to the most different states of the molecular system. To eliminate artificial combinations of maxima we associate to each combination a weighting factor being defined as the number of states that are within a predefined Euclidean distance to the maxima combination. We sort with respect to the weighting factors and neglect all combinations with zero weight; the representative conformations are then defined as the states that are closest to the combination of maxima.

CLUSTERING METHODS

In this section we present two different concepts to identify essential molecular conformations. Although they are based on different aspects of the simulation data, both methods are grounded on the same idea: they exploit special properties of eigenvectors corresponding to a so-called proximity matrix associated with the system.

To introduce the identification methods consider the following setting: given a set of data $\{d_1, \ldots, d_s\}$, for example, single configurations or sets of configurations, and a proximity function $p(d_i, d_j) \in [0, 1]$ measuring the degree of association between two elements. In the case of the methods presented below, the proximity function measures either structural or dynamical relations. It is $p(d_i, d_j) \approx 1$ for strongly related data and $p(d_i, d_j) \approx 0$, if d_i and d_j are only weakly related. [We do not request symmetry, i.e., $p(d_i, d_j) = p(d_j, d_i)$ for all i, j, for the proximity function.]

We are interested in decomposing the set of data into disjoint clusters (conformations) $C_1, \ldots, C_c$, such that each cluster C_i groups together related elements, while elements of different clusters are mostly unrelated. Let $p(C_i, C_j)$ denote the proximity between the two clusters C_i, C_j defined by an appropriate average value of $p(d, \hat{d})$ for $d \in C_i$ and $\hat{d} \in C_j$. Then we ask for a decomposition into clusters $C_1, \ldots, C_c$, such that

$$p(C_i, C_i) \approx 1 \quad \text{and} \quad p(C_i, C_j) \approx 0, \qquad i \neq j. \tag{1}$$

To identify the clusters, both methods use a proximity matrix $P = (P_{ij})$ based on the proximity function p. The off-diagonal entries are given by $p(d_i, d_j)$, while the diagonal entries are different for either method.

In view of eq. (1), the clustering problem is equivalent to finding a permutation of the data $d_1, \ldots, d_s$ such that the permuted proximity matrix is as block diagonal as possible, in the sense that the average value over off-block diagonal entries is much less than the corresponding block diagonal value. Because the problem of finding such a permutation is in practice unsolvable (it is a so-called *NP*-complete problem), the methods presented below pursue (different) heuristics. However, both exploit eigenvectors of the proximity matrix.

STRUCTURE-BASED CONFORMATIONS

The structural method aims at classifying configurations according to their structural proximity or similarity. The set of configurations is partitioned into disjoint subsets with the property that two configurations in the same subset are in some sense structurally closer to each other than two configurations in different subsets. In the statistical literature, this problem is known as cluster analysis, i.e., the classification of a set of feature vectors by their intrinsic properties.[9, 10, 12]

The measure of structural proximity should be invariant under rotations and translations. We thus choose to describe a configuration q not by the Cartesian positions of its atoms, but by its intramolecular distances defining a symmetric $n \times n$ distance matrix $D(q) = (D_{ij}(q))$ with

$$D_{ij}(q)^2 = \sum_{k=1}^{3} |q_{ik} - q_{jk}|^2, \qquad i,j = 1,\ldots,N.$$

[For the sake of simplicity, we assume the components of a configuration $q = (q_1,\ldots,q_{3N})$ ordered in the way that (q_{i1}, q_{i2}, q_{i3}) represents respectively the x,y,z positions of the ith atom.]

In the statistical literature, the distance matrix is known as feature vector. Because D is symmetric, it is sufficient to consider only the $N(N-1)/2$ different intramolecular distances. With respect to the number of degrees of freedom, the set of different intramolecular distances is still overdetermined, which allows to further reduce the distance matrix (see Results).

The structural distance between two configurations q and $\hat{q}$ is defined as the distance of their distance matrices:

$$\mathrm{dist}(q,\hat{q}) = \frac{1}{N(N-1)/2}\left(\sum_{i<j}^{N} \left(D_{ij}(q) - D_{ij}(\hat{q})\right)^2\right)^{1/2}.$$

To transform the structural distance into a proximity, we choose the function

$$p(q,\hat{q}) = \exp\left(-\frac{\mathrm{dist}(q,\hat{q})}{c}\right)$$

with a suitably chosen constant c, which sets the preferred distance scale of the clusters. [Their are other possible transformations, for example, $p(q,\hat{q}) = 1 - \mathrm{dist}(q,\hat{q})/c$ or $p(q,\hat{q}) = 1/(1 + \mathrm{dist}(q,\hat{q})/c)$. However, our cluster algorithms perform best with the proximity function given above.] Distances much smaller than c are mapped to nearly complete similarity, and will almost always form a cluster, while distances much larger than c are mapped to nearly complete dissimilarity and will rarely form a cluster. We found that good values of c are between 1/10 and 1/100 of the maximum distance of any two configurations in the data set. The exact value depends on the application, whether a rough or accurate seperation of clusters should be performed. The proximity between two clusters C_i and C_j is then defined as

$$p(C_i, C_j) = \sum_{q\in C_i,\ \hat{q}\in C_j} p(q,\hat{q}). \tag{2}$$

The identification of clusters or conformations is attacked hierarchically by gradually partitioning a cluster C into two subclusters C_+, C_-. Its basic idea is due to:[11, 22, 23] let χ_i be $+1$ if configuration $q^{(i)}$ belongs to C_+, and -1 if it belongs to C_-. Then

$$\begin{aligned} p(C_+,C_-) &= \frac{1}{8}\sum_{i,j=1}^{S} p(q^{(i)},q^{(j)})(\chi_i - \chi_j)^2 \\ &= \frac{1}{4}\sum_{i,j=1}^{S} p(q^{(i)},q^{(j)})\chi_i^2 - \frac{1}{4}\sum_{i,j=1}^{S} p(q^{(i)},q^{(j)})\chi_i\chi_j \\ &= \frac{1}{4}\sum_{\substack{i,j=1\\ i\neq j}}^{S} \chi_i\, p(q^{(i)},q^{(j)})\chi_i - \frac{1}{4}\sum_{\substack{i,j=1\\ i\neq j}}^{S} \chi_i\, p(q^{(i)},q^{(j)})\chi_j \\ &= \frac{1}{4}\sum_{i,j=1}^{S} \chi_i P_{ij}\chi_j \;=\; \frac{1}{4}(\chi, P\chi), \end{aligned}$$

where $(\cdot,\cdot)$ denotes the Euclidean inner product and P the proximity or Laplacian matrix:

$$P_{ij} = \begin{cases} -p(q^{(i)},q^{(j)}) & \text{if } i \neq j, \\ \sum\limits_{\substack{k=1\\ k\neq i}}^{S} p(q^{(i)},q^{(k)}) & \text{if } i = j. \end{cases}$$

The cluster problem is to minimize $(\chi, P\chi)$ under the constraint that $\chi_i = \pm 1$ and not all χ_i are equal. A similar minimality condition arises in diagonalizing a symmetric matrix: the eigenvector X corresponding to the lowest eigenvalue of the matrix minimizes (X, PX) under the constraint that X is normalized; the eigenvalue is equal to the value attained at the minimum. Higher eigenvectors minimize this quantity under the additional constraint of being orthogonal to all lower eigenvectors.

By construction, the matrix P has a trivial lowest eigenvalue of zero corresponding to the constant eigenvector $(1,\ldots,1)$. The second lowest eigenvector X then represents a minimum of (X, PX) where X is orthogonal to the constant vector, i.e., has vanishing sum. It, thus, will consist of positive and negative (and possibly zero) entries. This differs from the original constraint only in allowing noninteger (or zero) values. We can consider the second eigenvector as a heuristic approximation to the true solution of the discrete problem by mapping "continuous" entries X_i to discrete values $\chi_i = \pm 1$ using a threshold t:

$$\chi_i = \begin{cases} -1 & \text{if } X_i < t, \\ +1 & \text{if } X_i \geq t. \end{cases}$$

The threshold t is chosen automatically in each subclustering step: (1) choose $t = \max(X_i)$, (2) identify the two subsets $C_+ = \{i\colon X_i \geq t\}$ and $C_- = \{i\colon X_i < t\}$, (3) determine $P(C_+, C_-)$, and (4) repeat (2) + (3) for decreasing t-values until $t = \min(X_i)$, and choose as a partition the two subsets $C_\pm$, for which $P(C_+, C_-)$ is minimal. Typically the threshold t will be close to zero. In this way, the second-lowest eigenvector serves as a heuristic for determining the partitioning that must be considered in the minimization procedure. Instead of all partitions, the algorithm considers only those arising from thresholding the second eigenvector. This splitting procedure is repeated iteratively.

Because the proximity of two clusters as defined in eq. (2) is not weighted by the size of the subclusters, it favors splitting off a subcluster containing only a single configuration. (In this case, there will only be $S - 1$ terms in (2), as opposed to $S^2/4$ for a symmetric split.) One could include a suitable weighting factor in the definition of proximity, but this would destroy the connection to the eigenproblem of the Laplacian matrix, although it is still possible to use such a weighting factor when the threshold t is determined. However, splitting off single configurations only indicates that these configurations do not cluster very well. We found that after a few of these peripheral splits, there occur central splits that result in subclusters of a comparable size.

DYNAMICS-BASED CONFORMATIONS

The dynamics-based clustering characterizes conformations in terms of *meta*-stability. The state space is partitioned into disjoint subsets (conformations) with the property that each subset is *meta*-stable with respect to fluctuations within the canonical ensemble. Consequently, transitions between different subsets are rare events. In the first part, we introduce the dynamics within the ensemble and how to approximate it; in the second part, we present a cluster method to identify *meta*-stable conformations.

The distribution of molecular systems within the canonical ensemble does not change in time. However, there are fluctuations within the ensemble, because single systems evolve according to the Hamiltonian equation of motion. To capture the internal dynamics, we fix a time τ and observe all single system fluctuations after the span τ. A subset $C \subset \Omega$ is called invariant under the dynamics, if all systems being in C stay there after the time τ. By definition, the whole configurational space is invariant. Furthermore, a subset C is called almost invariant or *meta*-stable, if most of the systems stay in C after time τ.

To introduce the measure of dynamical proximity and, therefore, make the above characterization more precise, we approximate the fluctuations within the canonical ensemble. Based on the simulation data $q^{(1)}, \ldots, q^{(S)}$, we choose for each configuration $q^{(k)}$ a momentum $p^{(k)}$ according to the density P (see above) and integrate the Hamiltonian system for the time τ. This results in new configurations $\hat{q}^{(k)} = q(\tau; q^{(k)}, p^{(k)})$ and transitions $q^{(k)} \to \hat{q}^{(k)}$. (For better approximation results, choose momenta $p^{(k)_1}, \ldots, p^{(k)_m}$ according to P. This results in m transitions $q^{(k)} \to \hat{q}^{(k)_1}, \ldots, q^{(k)} \to \hat{q}^{(k)_m}$.)

For two subsets $C_{\text{from}}, C_{\text{to}}$, the dynamical proximity $p_\tau(C_{\text{from}}, C_{\text{to}})$ measures the relative frequency of transitions from C_{from} to C_{to}:

$$p_\tau(C_{\text{from}}, C_{\text{to}}) = \frac{\text{no.}(q^{(k)} \in C_{\text{from}} \text{ and } \hat{q}^{(k)} \in C_{\text{to}})}{\text{no.}(q^{(k)} \in C_{\text{from}})}; \quad (3)$$

it can be interpreted as the conditional transition probability of being in C_{from} and changing to C_{to} within the time τ. We call a subset or conformation C *meta*-stable, if $p_\tau(C, C) \approx 1$ and a transition a rare event, if $p_\tau(C_{\text{from}}, C_{\text{to}}) \approx 0$. In contrast to the structure-based method, the proximity is defined only for subsets of states rather than for single states. To solve the dynamics based cluster problem, i.e., to identify *meta*-stable conformations, we, therefore, discretize the state space into disjoint sets $B_1, \ldots, B_d$, for example, boxes resulting from a grid defined by partitioning the Cartesian coordinates or torsion angles into intervals, and seek out clusters that can be written in terms of these sets B_i. At this point, we are in the situation of the general cluster problem presented at the beginning of this section. Thus, in principle, every cluster algorithm based on the discretization sets $B_1, \ldots, B_d$ and the transition probabilities $p_\tau(B_i, B_j)$ could be applied to identify *meta*-stable conformations. In the following, we present a cluster method that exploits the special structure of the transition probabilities and can be interpreted as the discretization of a continuous cluster problem (for details, see ref. 16).

As for the structural method, the identification algorithm is based on a proximity or transition matrix $P = (P_{ij})$, which is defined by the transition probabilities, $P_{ij} = p_\tau(B_i, B_j)$. For the identification process, we exploit the following two properties of the transition matrix (for more details see ref. 24): (1) The transition matrix is stochastic, i.e., its entries are nonnegative, and the sum of each row equals one. As a consequence, the constant vec-

tor $(1, \ldots, 1)$ is an eigenvector corresponding to the eigenvalue $\lambda_1 = 1$. (2) The presence of *meta*-stable conformations corresponds to a block structure of the transition matrix (for a suitable permutation of the B_i) and a splitting of the spectrum into a cluster of eigenvalues $\lambda_1, \ldots, \lambda_c$ near 1 and the remaining part of the spectrum. The two spectral parts are separated by a gap. The number of *meta*-stable conformations, blocks in the transition matrix and eigenvalues near 1 are equal.

It follows from perturbation analysis[24] that the eigenvectors $X_1, \ldots, X_c$ corresponding to the cluster of eigenvalues near 1 are almost constant on each *meta*-stable conformation, i.e., if B_i and B_j belong to the same conformation, then $X_k(B_i) \approx X_k(B_j)$ for $k = 1, \ldots, c$. Furthermore, the c-tuple of eigenvector components associated with each B_i,

$$B_i \mapsto \big(X_1(B_i), \ldots, X_c(B_i)\big),$$

is sufficient to identify the conformations in the case of weak coupling.[24] Each conformation is the collection of sets B_i with an almost identical c-tuple. Thus, using the eigenvectors $X_1, \ldots, X_c$, we have incorporated the dynamics by encoding the discretization sets $B_1, \ldots, B_d$ through c-tuples. The identification of conformations is reduced to clustering these c-tuples with respect to (geometrical) similarity. We have implemented an algorithm that also copes with larger perturbations in the eigenvector components due to stronger coupling between the conformations. A detailed description of the algorithm is given in ref. 24.

At the end of this section, we want to address the problem of how to choose the discretization sets $B_1, \ldots, B_d$. On the one hand, because we seek out dynamical conformations as unions of the B_i, the partitioning should be as fine as possible in order to allow "arbitrary-shaped" conformations. On the other hand, a fine partitioning requires many states and transitions to accurately determine the transition probabilities in (3). We adopt the following strategy: because conformational transitions should correspond to changes in the essential degrees of freedom, we define the discretization sets only in terms of these essential coordinates (see Representative Conformations). Furthermore, we take only those essential degrees of freedom into account whose distributions are far from being broad Gaussian shaped. The distributions are partitioned into their Gaussian-like parts, and the B_i are defined as collection of states that fit into a certain combination of these subdivisions.

Results

The approaches to identify representative, structure and dynamics-based conformations were applied to the triribonucleotide adenylyl(3′-5′) cytidylyl(3′-5′)cytidin [r(ACC)] model system in vacuum (Fig. 3). It consists of $N = 70$ atoms, whose physical representation is based on the GROMOS96 extended atom force field.[25]

SAMPLING OF THE CANONICAL DENSITY

The simulation data were generated by means of an ATHMC sampling of the canonical density at $T =$ 300 K. The subtrajectories of length 80 fs (femtoseconds) were computed by means of the Verlet discretization with a stepsize of 2 fs. For these parameters, HMC simulations typically require thousands of iterations only to leave the neighborhood of the initial configuration. Application of ATHMC (with adaptive temperatures between 300 and 400 K) circumvents the problem: one observes frequent transitions in the crucial torsion angles of the molecule (for details see ref. 18). The simulation was divided into four Markov chains, each starting with a different state chosen from a high temperature run at 500 K, which allowed the molecule to move into different conformations. The sampling took about 12 h on a workstation with MIPS R10,000 processor. It was terminated by a convergence indicator[26] associated with the potential energy and all 37 torsion angles after 320,000 steps, resulting in the sampling sequence $q^{(1)}, \ldots, q^{(S)}$, $S =$ 32,000 (considering only every 10th step). We have found slower convergence for the torsion angles of the terminal ribose. Because the temperature can change during the ATHMC run, each configuration is connected with a reweighting factor with respect to the canonical ensemble at 300 K.

REPRESENTATIVE CONFORMATIONS

For Cartesian coordinates and torsion angles there are, respectively, five and four essential degrees of freedom. The transformation process for the torsion angles is exemplified in Figures 1 and 2. Figure 1 (top, left) shows the circular deviations ρ of the transformed torsion angles in decreasing order of magnitude. Only the first four transformed torsion angles have relevant circular deviation and are far from being Gaussian shaped (see Fig. 2), while the remaining transformed torsion angles are Gaussian-like. (We scaled the transformed torsion angles such that the range is between -180 and

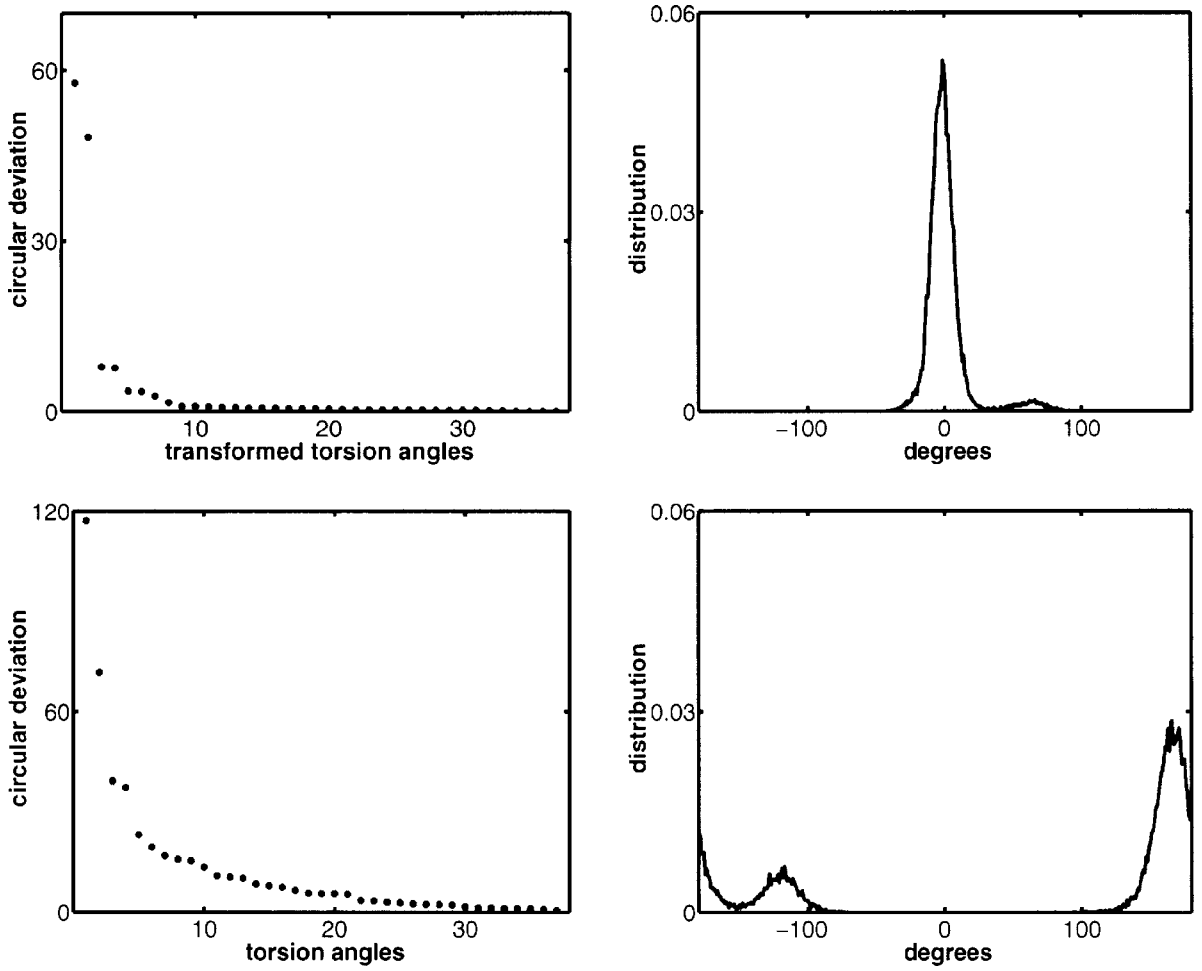

FIGURE 1. Top: circular deviation of the transformed torsion angles ordered by magnitude (left) and the distribution of the first nonessential torsion angle, which is the fifth angle in the sequence (right, see text). Bottom: circular deviation of the original torsion angles (left) and the distribution of the fifth torsion angle (right).

+180.) On the top right part of Figure 1 the first nonessential transformed torsion angle, which is the fifth in the sequence, is shown. The graphics at the bottom part of Figure 1 show the circular deviations of the original torsion angles (left) in decreasing order, and the distribution of the fifth torsion angle (right) in sequence. Clearly non-Gaussian distributions show about 10 of the 37 torsion angles.

To identify representative conformations (R-conformations), we determine the maxima for each distribution of the essential torsion angles (see Fig. 2). These maxima have been grouped to $3 \times 3 \times 2 \times 2 = 36$ combinations, each of them defining a theoretical, but not necessarily realized configuration of the molecule. From 15 maxima combinations with nonvanishing weight, we have selected two representative conformations to visualize characteristic differences (Fig. 3).

STRUCTURE-BASED CONFORMATIONS

Because for the current structural cluster algorithm the computational effort grows quadratically with the number of configurations, we have performed the algorithm based on a subensemble of 1000 configurations out of the 32,000 sampling configurations. (This selection is realized randomly, taking the different statistical weights of the configurations into account. Using a subensemble one is always in danger of loosing relevant information. Thus, it should be emphasized that the selection step is not necessary, i.e., one could also compute the required eigenvectors of the proximity matrix for the entire data set by applying subspace-oriented iterative eigenvalue solvers.[29, 30] However, in the case considered herein, the results to be presented do not depend sensitively on the length of the subensemble.)

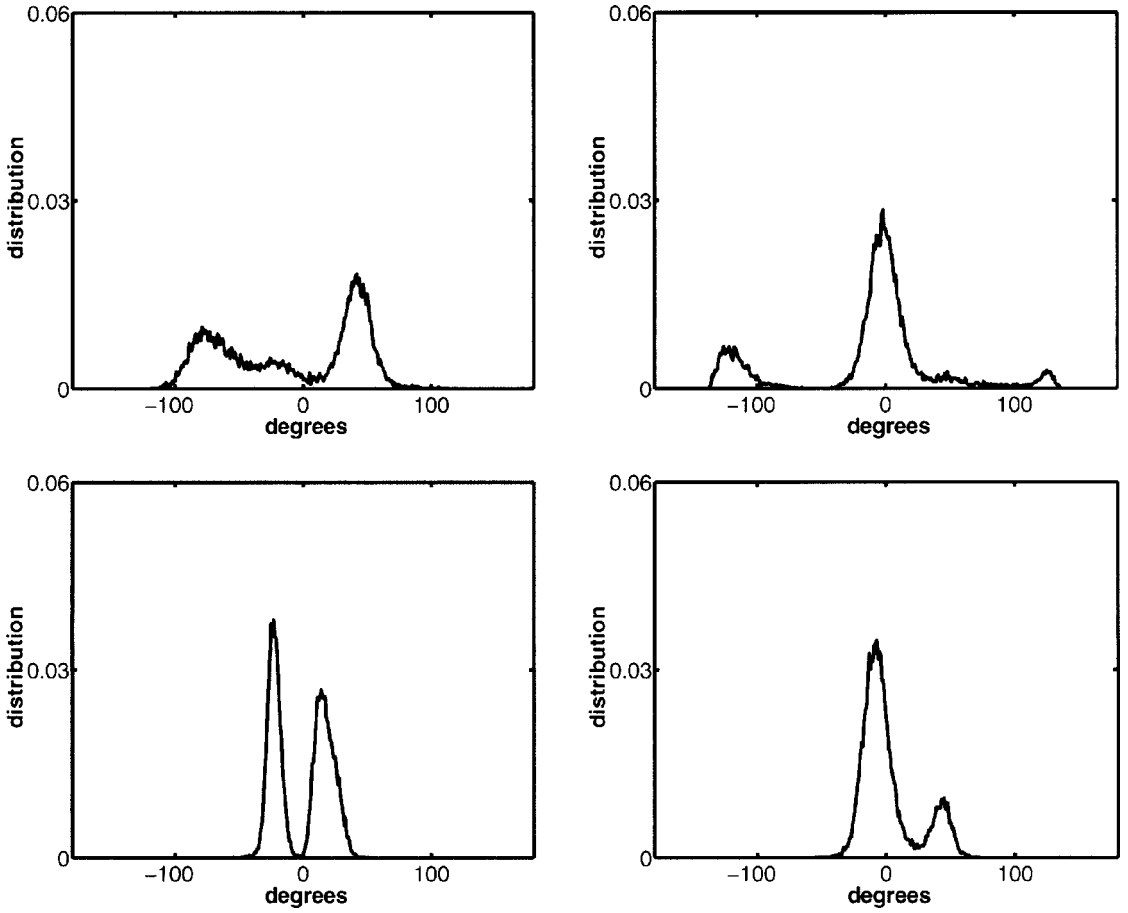

FIGURE 2. Distribution of the four essential torsion angles. The distributions at the top allow identification of three maxima each, while there are two maxima for each distribution at the bottom.

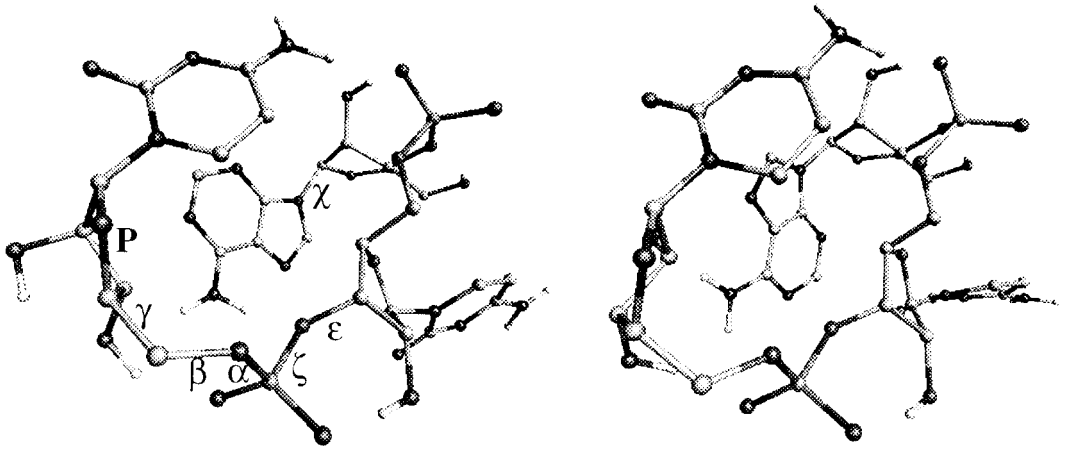

FIGURE 3. Two conformations of r(ACC). Left: The χ angle around the first glycosidic bond is in *anti* position (−175 degrees) and the terminal ribose pucker *P* is in C(3′)endo C(2′)exo conformation. Right: The χ angle is in *syn* position (19 degrees) and the terminal ribose in C(2′)endo C(3′)exo conformation.

As stated in the Structure-Based Conformations section, the set of intramolecular distances is overdetermined with respect to the number of degrees of freedom of a configuration. To emphasize differences between conformations and suppress smaller statistical vibrations, we reduce the set of

intramolecular distances to those with largest statistical variance in the sampling. To make sure that the reduced distance matrix still describes the whole molecule, we have selected $3N$ pairs of atoms with the largest variance such that each atom occurs three times in the pairs.

The identification of clusters is performed hierarchically. Clusters are split into two subclusters according to the measure of similarity (with c equal to 1/30 of the maximum distance of any two configurations in the sampling). During the process, many splits remove only a few data points from the parent cluster, i.e., the cluster from which the current cluster was split off, as our measure of proximity is inclined to do for combinatorial reasons. As a consequence, we consolidate the cluster hierarchy by removing all clusters that have fewer than 10 data points. We characterize the clusters by the average distance of their members in comparison to the average distance of all data points or to the average distance in the father cluster. The latter measure indicates, in particular, good clustering properties.

We use a multidimensional scaling plot (Figs. 4 and 5) to visualize clusters.[27] The $2D$ plot shows a $2D$ least-squares approximation of the $3N = 210$ dimensional configurational space in the sense that neighboring points correspond in general to structurally similar configurations, while distant points reflect, in general, structural differences.

The structural clusters (S-clusters) are shown in Figure 4. On the top level of hierarchy, there are three well-separated clusters S1&S2, S3&S4, and S5, two of which can be split again giving a total of five clusters S1, . . . ,S5. The structural clustering took less than 4% of the computing time required for generation of the simulation data.

DYNAMICS-BASED CONFORMATIONS

The dynamical fluctuations within the canonical ensemble were approximated by integrating four short trajectories of length $\tau = 80$ fs starting from each sampling point $q^{(1)}, \ldots, q^{(S)}$. To facilitate transitions, analogous to the ATHMC sampling, the momenta were chosen according to the momenta distribution $P(p)$ for four different temperatures between 300–400 K, and reweighted afterwards. This resulted in a total of $4 \times 32{,}000 = 128{,}000$ transitions. This calculation took less than 25% of the total computing time.

The configurational space was discretized into boxes $B_1, \ldots, B_d$, by means of all four essential degrees of freedom (see Fig. 2) resulting in $d = 36$ discretization boxes. Then the 36×36 transition matrix P was computed based on the 128,000 transitions taking the different weighting factors into account. Because every box had been hit by sufficiently many transitions, the statistical sampling was accepted to be reliable. The computation of the eigenvalues of P near 1 yielded a cluster of eight eigenvalues with a significant gap to the remaining part of the spectrum:

k	λ_k
1	1.000
2	0.999
3	0.989
4	0.974
5	0.963
6	0.946
7	0.933
8	0.904
9	0.805
⋮	⋮

Finally, the dynamics based conformations (D-conformations) were computed based on the corresponding eight eigenvectors of P via the cluster algorithm presented in the Dynamics-Based Conformations section. We found eight D-conformations, which we have displayed in the multidimensional scaling plot based on structural proximity (Fig. 5). The clustering turned out to be rather insensitive to further refinements of the discretization. The weighting factors within the canonical ensemble and the *meta*-stability $p_\tau(D, D)$ of the eight identified conformations are given in the following table:

Conformations	Weighting Factor	*Meta*-Stability
D1c	0.107	0.986
D1t	0.011	0.938
D2c	0.116	0.961
D2t	0.028	0.888
D3c	0.320	0.991
D3t	0.038	0.949
D4c	0.285	0.981
D4t	0.095	0.962

The transition probabilities between the different D-conformations are visualized schematically in Figure 6. In the limit of infinitely many transitions,

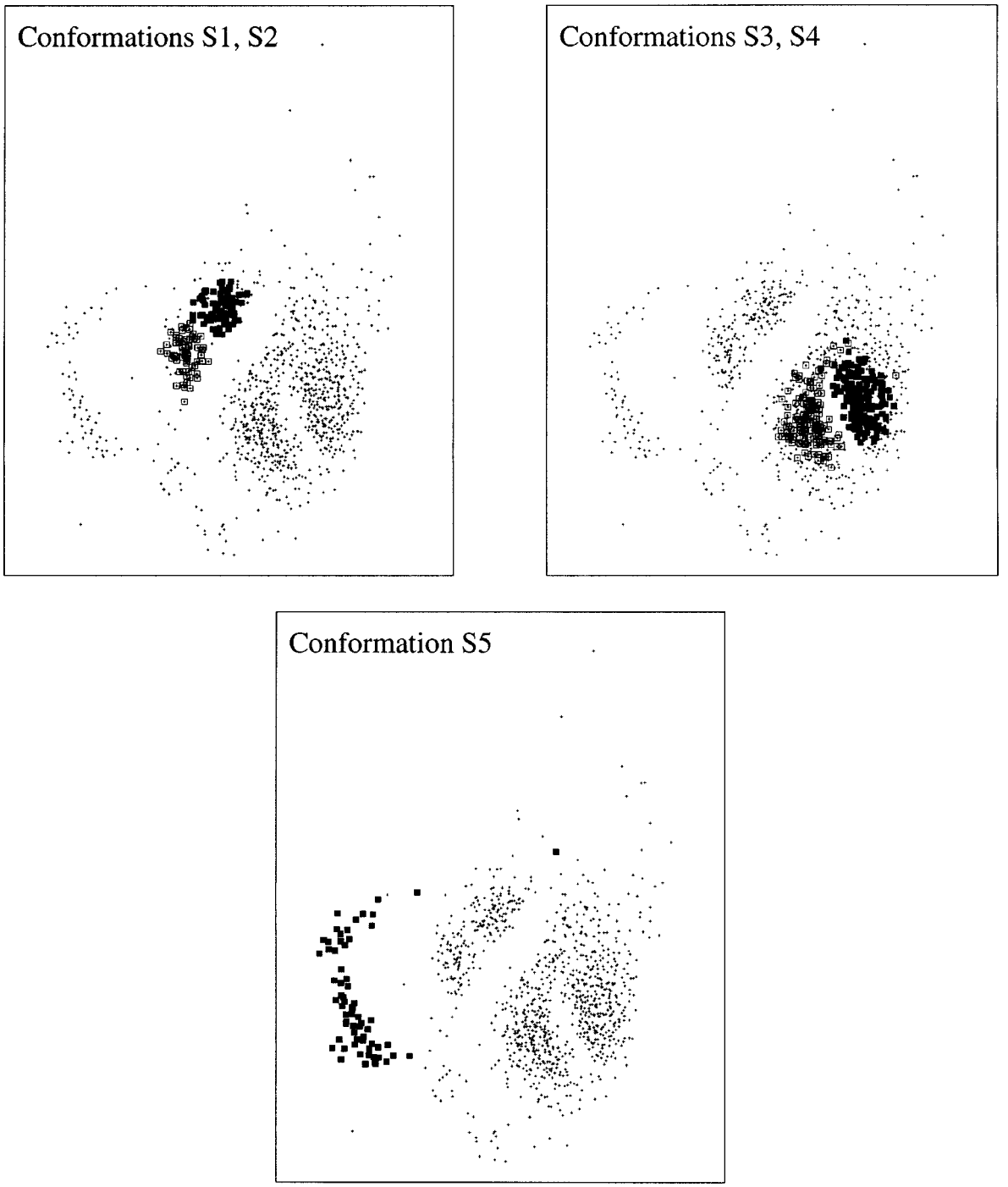

FIGURE 4. *2D* plot of the five structure-based conformations S1, . . . ,S5. The distinction between open and filled squares indicates a the splitting into subsets.

the transition matrix should exploit a reversible, but not symmetric structure. Furthermore, the matrix allows definition of a hierarchy between the clusters, which is inherent to the algorithm. On the top level, there are two clusters, D1&D2 and D3&D4, corresponding to the two 4×4 blocks on the diagonal. On the next level, each of these clusters split up into two subclusters yielding D1, . . . ,D4. On the bottom level, each cluster is further divided into a core (c) and a transition (t) part. The dynamical clustering took less than 2% of the comput-

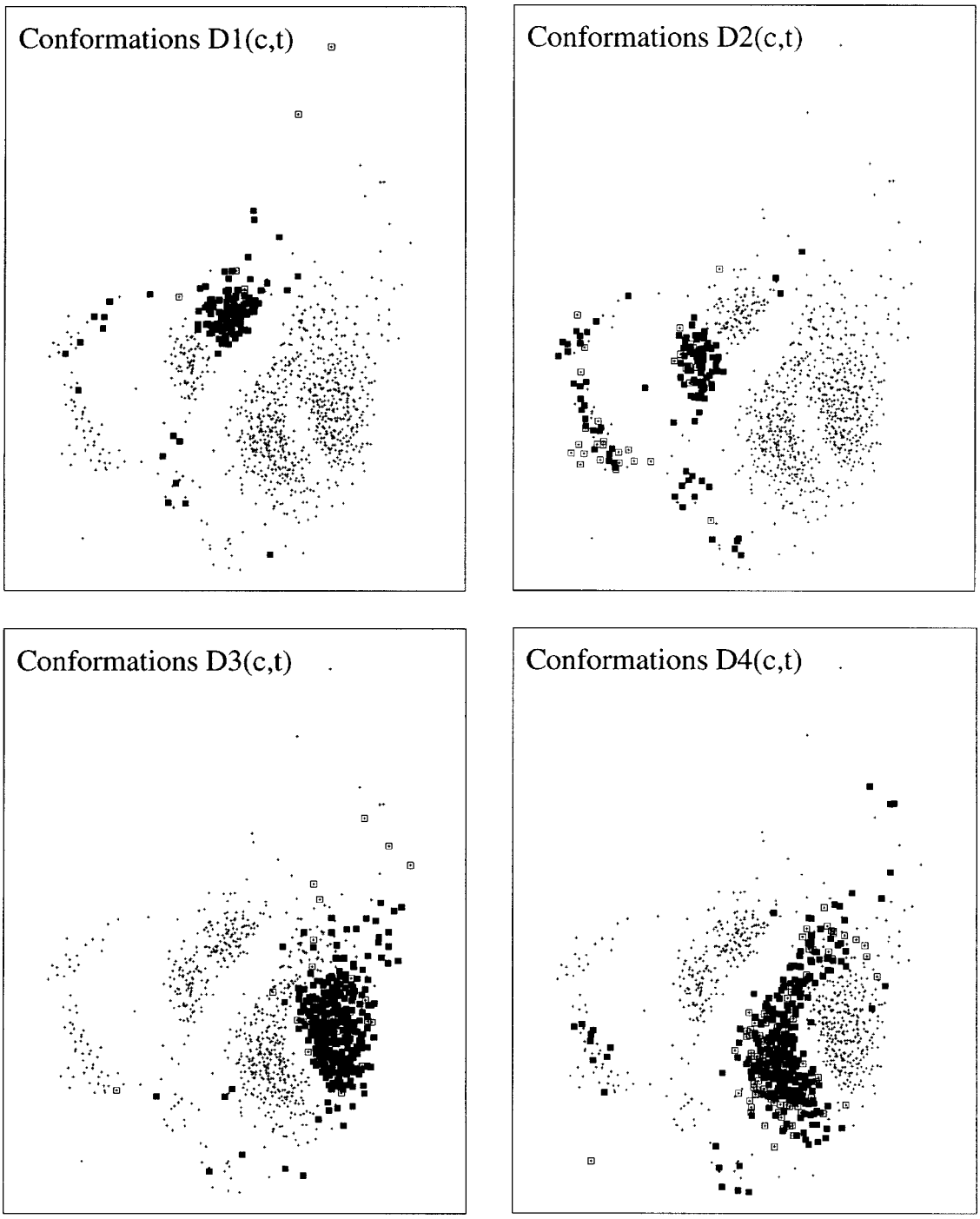

FIGURE 5. *2D* plot of the four dynamical conformations D1, . . . ,D4 (squares). The distinction between open and filled squares indicates a further splitting into eight conformations resulting from a partition into a core (c) and a transition (t) conformation.

ing time required for evaluation of the simulation data.

Discussion

Comparing dynamics and structure based conformations by means of Figures 4 and 5 manifests

the characteristics of each concept. The four confor-

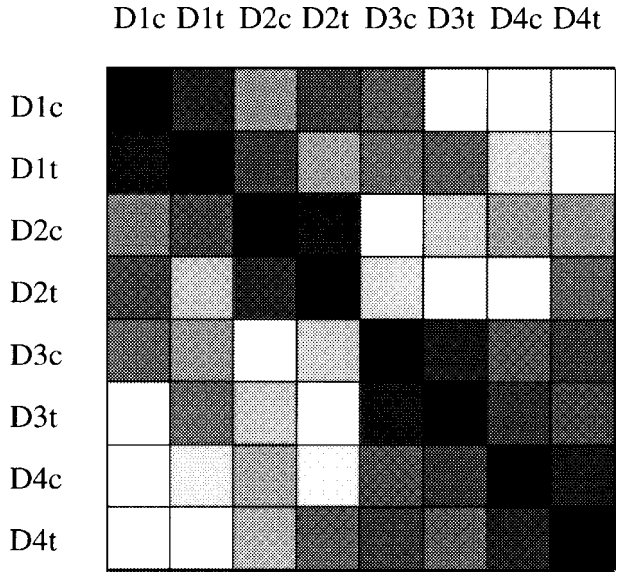

FIGURE 6. Schematical visualization of the transition probabilities $p_\tau(D_{\text{from}}, D_{\text{to}})$ between the dynamical conformation D_{from} (row) and D_{to} (column). The colors are chosen according to the logarithm of the corresponding entries; black: $p_\tau \approx 1$, white: $p_\tau \approx 0$.

mations S1,...,S4 correspond to D1,...,D4, while S5 is part of all four dynamical conformations. This is made more precise by the following table, which shows the percentage of S-conformations in D-conformations:

	D1	D2	D3	D4
S1	1.000	0.000	0.000	0.000
S2	0.000	1.000	0.000	0.000
S3	0.000	0.000	0.994	0.006
S4	0.000	0.000	0.000	1.000
S5	0.121	0.697	0.015	0.167

The table indicates that structural similarity and *meta*-stability might coincide. However, a difference between the two concepts is given by the last row. While S5 is composed of structurally similar configurations, it is by no means *meta*-stable, being spread out all over the four D-conformations. This might indicate that S5 is a collection of transition states that do not cluster due to their similarity among each other, but rather due to their structural difference to all other states. In the dynamical analysis, these transition states are grouped to those clusters to which they fluctuate most likely.

Figure 3 shows two representative conformations with characteristic differences, selected out of the 15 representative conformations with nonvanishing weighting factor (see the Representative Conformations section). Analyzing representative conformations in comparison to conformational ensembles,

we have found, that the first six conformations with highest weights belong to the four dominant dynamical clusters D1, . . . ,D4. Of course, with the described method we will expect to find more representative conformations than dynamical clusters. The combination of maxima in the distribution of transformed torsion angles actually indicates a possible discretization of the state space into relatively few boxes. Each *meta*-stable conformational ensemble is composed of some of these boxes such that the number of conformational ensembles should be less than (or at most equal to) the number of representative conformations.

The conformations in Figure 3 belong to the clusters D2 and D3, and can be visually distinguished according to the orientation of the χ angle around the first glycosidic bond and the conformation of the terminal ribose, indicated by the so called sugar pucker *P*. In order to analyze whether the two R-conformations allow representation of the two dynamical conformations, we plotted the distributions of the χ and one of the torsion angles (within the sugar pucker *P*) for the D2 and D3 conformation in comparison to all sampled states (Fig. 7). Obviously, the structural differences between the states in the conformational ensembles D2 and D3 can be described by differences in the ribose conformation and in the orientation of the adenine.

Torsion angle fluctuations at terminal groups have only minor influence on the global structure of the molecule. However, they may influence the covariance analysis based on torsion angles and, thus, the essential degrees of freedom. There are two possible remedies. On the one hand, one can switch to a representation in Cartesian coordinates, but then the overall translational and rotational motion of the molecule have to be eliminated. On the other hand, one can exclude all torsion angles corresponding to terminal groups for the covariance analysis. We favor the latter approach, because the chemical intuition and the imagination of global changes are facilitated in the space of torsion angles.

Concluding Remarks

It is intriguing to apply "data mining"-oriented statistical techniques to large data sets originating from MD or MC samplings of molecular ensembles. The idea of using cluster methods naturally leads to the concept of decomposing the data set into conformational ensembles or subsets, which can either be defined via their significant *meta*-stability or via structural similarities between the molecular configurations contained. This clustering of configurations into conformational ensembles is clearly different from the concept of choosing single conformations as representatives for common properties of a larger set of configurations. A few representative conformations allow a rough but fast examination of the state space, while the concept of conformational ensembles enable the investigation of structural conservation or *meta*-stability in the case of dynamically based methods. Additionally, the latter approach can be used to calculate transition rates between conformational ensembles and to locate transition states.

References

1. Frauenfelder, H.; Sligar, S.; Wolynes, P. Science 1991, 254, 1598.
2. Gerstein, M.; Lesk, A.; Clothia, C. Biochemistry 1994, 33, 6739.
3. Hayward, S.; Berendsen, H. Proteins 1998, 30, 144.
4. Zhou, H.; Wlodek, S.; McCammon, J. Proc Nat Acad Sci USA 1998, 95, 9280.
5. Grubmüller, H.; Heymann, B.; Tavan, P. Science 1996, 271, 997.
6. Duan, Y.; Kollman, P. Science 1998, 282, 740.
7. Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W.; Mark, A. J Mol Biol 1998, 280, 925.
8. Cordes, F.; Starikov, E.; Saenger, W. J Am Chem Soc 1995, 117, 10365.
9. Anderberg, M. R. Cluster Analysis for Applications; Academic Press: New York, 1973.
10. Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data; Advanced Reference Series ed.; Prentice Hall: Englewood Cliffs, NJ, 1988.
11. Hendrickson, B.; Leland, R. SIAM J Sci Comput 1995, 16, 452.
12. Drineas, P.; Frieze, A.; Kannan, R.; Vempala, S.; Vinay, V. Preprint; Yale University, Dept. of Computer Science, to appear in the Proceedings of the Symposium on Discrete Algorithms, SIAM (unpublished).
13. Kloppenburg, M.; Tavan, P. Phys Rev E 1997, 55, 2089.
14. Karpen, M.; Tobias, D.; Brooks, C. L. III, Biochemistry 1993, 32, 412.
15. Gordon, H.; Somorjai, R. Proteins 1992, 14, 249.
16. Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. J Comput Phys (Special Issue on Computational Biophysics) 1999, 151, 146.
17. Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Proteins 1993, 17, 412.
18. Fischer, A.; Cordes, F.; Schütte, C. J Comput Chem 1998, 19, 1689.
19. Hayward, S.; Kitao, A.; Berendsen, H. Proteins 1997, 27, 425.
20. Fisher, N. I. Statistical Analysis of Circular Data; University Press: Cambridge, 1993.
21. Fisher, N. I.; Lee, A. J. Biometrica 1983, 70, 327.

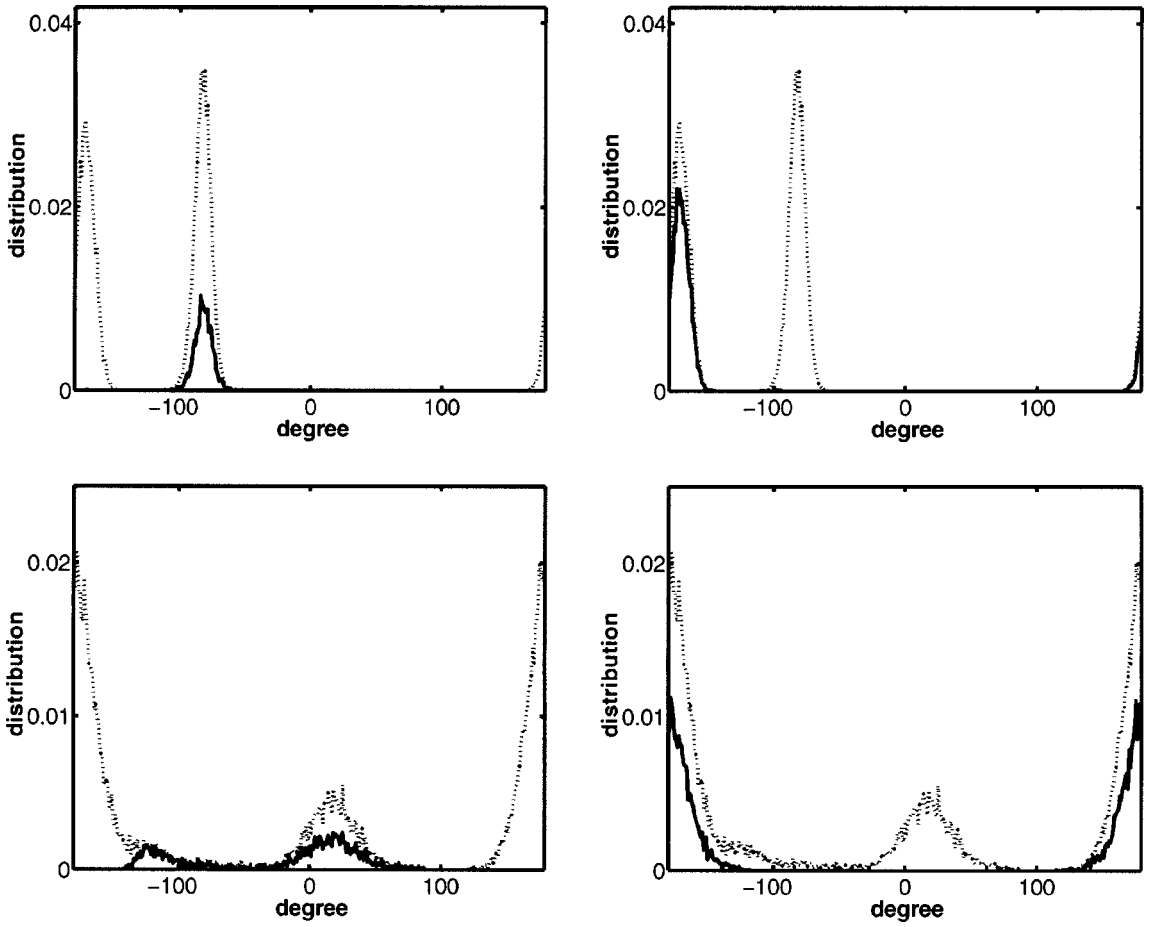

FIGURE 7. Distribution of torsion angles for all sampling data (dashed line) and selected clusters (solid line). Top: distribution of one torsion angle at the terminal ribose for the conformations D2 (left) and D3 (right). Bottom: distribution of the χ angle (see Fig. 3) at the adenine for the conformations D2 (left) and D3 (right).

22. Fiedler, M. Czech Math J 1975, 25, 607.
23. Fiedler, M. Czech Math J 1975, 25, 619.
24. Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Preprint SC-98-03, Konrad-Zuse-Zentrum, Berlin. Available via http://www.zib.de/huisinga (unpublished).
25. van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. Biomolecular Simulation: The GROMOS96 Manual and User Guide; vdf Hochschulverlag AG: ETH Zürich, 1996.
26. Gelman, A.; Rubin, D. Stat Sci 1992, 7, 457.
27. Best, C.; Hege, H.-C. Preprint SC-98-42, Konrad-Zuse-Zentrum, Berlin. Available via http://www.zib.de/MDGroup (unpublished).
28. Jupp, P. E.; Mardia, K. V. Int Stat Rev 1989, 57, 261.
29. Deuflhard, P.; Friese, T.; Schmidt, F. Preprint SC-97-55, Konrad-Zuse-Zentrum, Berlin. Available via http://www.zib.de/bib/pub/pw/ (unpublished).
30. Lehoucq, R. B.; Sorensen, D. C.; Yang, C. ARPACK User's Guide: Solution of Large Eigenvalue Problems by Implicit Restartet Arnoldi Methods; Rice University: Houston, TX, 1998.